nlp终极目标是“语言理解”，我个人想分为以下三个层次：

1. 句子特征简单提取和分类：典型例子是分词、命名实体识别，这类问题已经能比较好的解决，甚至用不上之前提到的没有提取 几个字/句子局部 的信息
2. 整个句子的转化与特征提取：典型例子是机器翻译、文本分类，这类任务需要对句子整体的特征进行提取，也是目前受到最多关注的方向
3. 基于句子、文本特征进行拓展：典型例子是对话系统、推荐系统，这类任务难度更大，需要匹配特定任务需求，不像分类2相对简单的对应(翻译对应、文本分类标签)

学术上主要是模型的创新，attention、seq2seq这些都是通用的，只是不同领域任务可能有不同需求或trick。接下来与其说认死一个应用方向，不如先主要关注分类2这个层次的发展，再根据学习之后的理解关注更小的方向或分类3的应用

lecture 10-11 neural machine translation and attention

在基本的seq2seq模型上的补充：

1. 常见的做法是把encoder的最后一个时刻作为decoder的第一层，必须用LSTM保持记忆，也可以将encoder最后一层输给decoder的每一层。

2. encoder最后一个隐藏层状态是全文的总结，往往在隐层单元特别是最后一个做一些扩展，如使用LSTM、GRU等

3.NMT的优势：

①End-to-end training优化同一个损失函数调整所有参数

②Distributed representation share strength更好地利用词语、短语之间的相似性

③Better exploitation of context利用更多上下文——原文和部分译文的上下文 生成的文本更加流畅 传统的4-gram 5-gram不现实

弱点：

无法显式利用指代相消、子句连接的结果，难以解释和控制，无常识、无法翻译有奇怪的输出、词典外的词、训练测试词域对应错误、较长文本、出现少的搭配...

4.评判标准bleu：

n-gram评判：3或4

翻译过短进行惩罚

并不完美：句子翻译非唯一答案、n-gram并不全面

https://images2017.cnblogs.com/blog/626346/201710/626346-20171030145636246-148630779.png

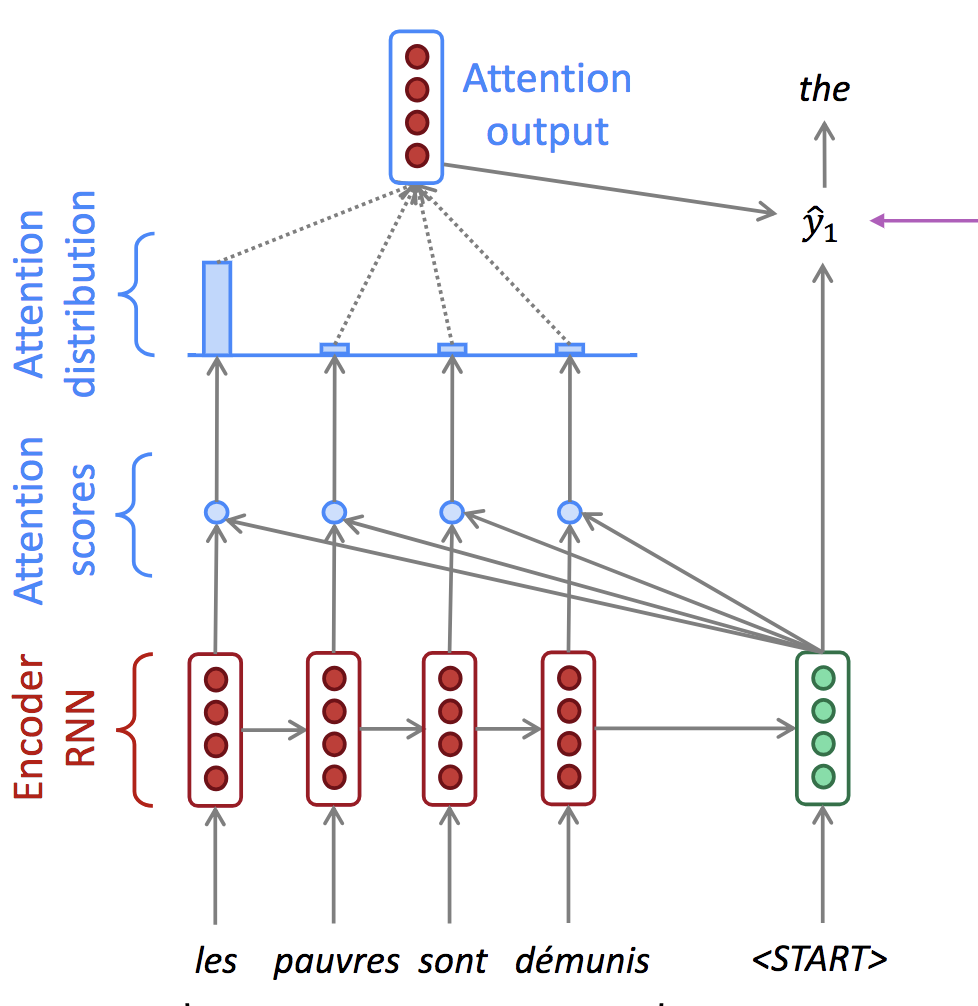
BP为长短惩罚因子，w为线性加权，Pn为对应n-gram的准确度

5.attention模型

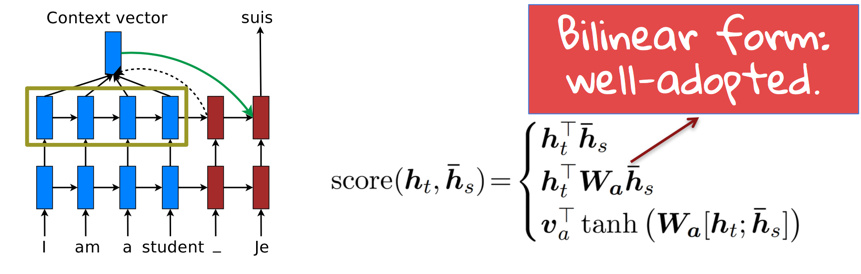
①attention相当于延长了记忆的持续时间，encoder结束相当于扫一遍句子， 通过attention的特定访问调取翻译词

②attention优点：提高准确率、对解决梯度消失有帮助、同时学习翻译和对齐

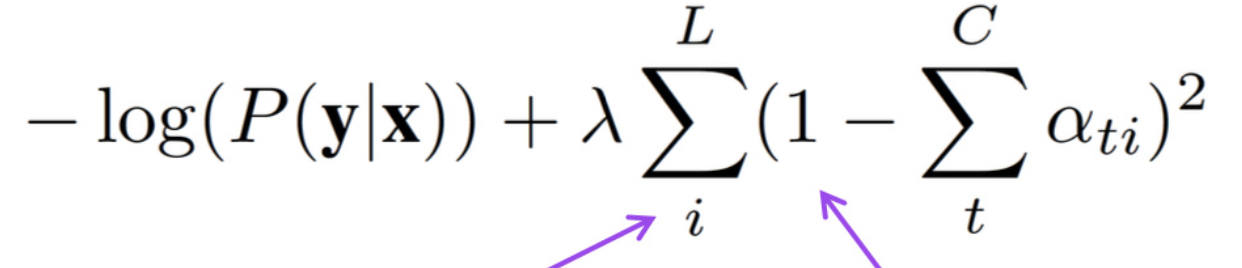
③基础模型：



打分方式：输入词的隐层和输出层的隐层



一些变式：local-attention、double-attent(同时注意原文和译文)

损失函数优化：防止一个词翻译太多词

④解决大词表(softmax)问题：

分层softmax或负采样

Large-vocab NMT：每次只在词表的一个小子集上训练

训练trick：让用词相似的文章进入同一个子集，这样每个子集的词表就变小了。

测试trick：将前K个最常使用的单词加入备选词表，然后将原文中每个单词可能的前K′个翻译加进去。最后在备选词表上softmax。

⑤还有更多更多的trick...更多更多更多的论文...to be continued...

这几个课后材料主要是经典模型的论文，只记录一些之前没注意到的：

Bleu:a method for automatic evaluation of machine translation

思路：

1.unigram需要剔除正确答案中已匹配的

修改n-gram模型基于 adequcy相同词语匹配 fluency是longer n-gram matches

2.发现4-gram的区分度最高

3.采用具有归一化系数的平均对数

4.惩罚在reference中没出现的词，惩罚候选中出现次数超出在reference中次数的词

Seq2seq

1.把输入的Seq按照相反的顺序输入，其性能表现还更好，猜测原因是目标句子与源句子开头的短时联系更加紧密了

2.可以学到对词序敏感的特征和翻译时主动被动形式的特征

3.encoder和decoder的LSTM是两个不同的模型，参数也不共享

Massive Exploration of Neural Machine Translation Architectures

NMT大量的实验和对参数的一些探索

1.encoders比decoders难训练，beam search很重要，LSTM比GRU好

2.word embedding中映射到小维度和大维度的梯度更新没有明显差别，梯度更新范数大致不变，大维度并不会导致过拟合，对效果影响很小

3.encoder层数为2、decoder层数为4时效果较好

4.attention模型梯度更新更大，文章中称为”weighted skip connection”,优化了梯度流

下周继续根据网课给的机器翻译 attention相关的论文进行阅读